



***A Study of Multivariate Behavior and Anomaly  
Patterns:  
Tensor Decomposition for Multiway Big Data***

A Dissertation submitted for

The partial fulfillment of

**Master of Engineering Research**

By

**Alina Rakhi Ajayan**

**University of Technology Sydney**

**New South Wales, Australia**

**2017**

## **CERTIFICATE OF ORIGINAL AUTHORSHIP**

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and fully acknowledged within the text.

I also certify that I have written the thesis. Any help that I have received in my research work and the preparation of the dissertation itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student:

Date:

## ACKNOWLEDGEMENT

My most profound gratitude is expressed to the University of Technology Sydney, and the School of Electrical and Data Engineering, FEIT, for this valuable opportunity to pursue impactful research. My attempts and motivation were primarily fuelled by my parents Mrs. Aleykutty Robert, Mr. G. R. Ajayan and my dear brother Akhil Minu Ajayan. I wish to express my gratitude to Dr. Zenon Chaczko, my advisor and supervisor through this two-year academic period.

The real drive and support came from my colleagues and friends at UTS – Manisha, Augustin, Hayat, Daniel, Firas, Ammar, Maria, Alaa, Mahmoud, Mohammad, Asma, Deepak, and Saha had been the constant voices in my head, egging me onward towards my goals and pursue my dreams. The entire Centre for Health Technologies Social group, starting with Dr. Adrian Bishop and Prof. Gyorgy Hutvagner had always been good mentors and advisors.

My candidature assessment panel members – the Hon’ble Chair and my Examiners, made my progress take a giant leap from the state of an in-depth Literature Review to a narrowed-down, streamlined research project. I hence extend my heartfelt gratitude to Dr. Daniel Franklin, Prof. Robin Braun, Dr. Negin Shariati Moghadam and Dr. Yi Cao, who brought me forward to the completion of this work.

Finally, but most importantly, the last few months of perseverance and dedication was hard to fight for and succeed. I could only push myself through the harsh write-up and implementation phase of research by the guidance of A/Prof. Raghav Menon and Michal Skowron.

Thanking you all from the bottom of my heart, for encouraging me to do my best, and to be at my best.

## ABSTRACT

A vast majority of the today's information haul is through Cyber-Physical Systems (CPS). They represent the confluence of extensive data sets, tight time-constraints, latency issues and heterogeneous components. CPS architectures demand newer Big Data processing approaches. Typical applications span from the Internet-of-Things, across the World Wide Web to Smart Cities and Intelligent machines. A standard heterogeneous CPS installation, the Smart Energy Grid, is observed and the logistics are analyzed. The Smart Grid domain is weighed down by lack of unifying framework and systemic intelligence for autonomic management. Preliminary studies of the field under investigation shows how processing of Real-Time data, communication and control signaling is vital. Purely autonomic system governance is proven to be different from the contemporary definition. It takes the form of Interoperability (achieved through automation) instead of elemental Integration. That means autonomic (smart) management requires all elements to have fully controllable behavior.

This dissertation tests the hypothesis of applying Tensor decompositions and Factorizations - a momentum-gaining arithmetic tool - to this problem. The aim is to validate the prospects of higher order Anomaly Pattern Processing to capture intelligence along multiple modes of data flow. Tensorial Data representation captures information flows in Big Data, while Multivariate Anomaly Detection performs tracking of the time-series behavioral changes. Together, they implement Autonomic management in CPS super-systems. Uniqueness of this approach is highlighted by the novel multi-modal data flow imaging and models. Requirements of traditional anomalous event definition and cataloging in Data streams are removed. Tensor algebra is then studied for the scope of implementation concerning features, significance, and interpretation in terms of multi-modal data. Standard Decomposition rules and their derivatives, literature analysis on contemporary applications of Tensor algebra, and its scope on prominent real-world data processing problems are studied. Finally, the decomposition tool for Multi-way analysis is inferred, and proposed methodology is derived.

The Smart Grid Smart City Project commissioned by the Australian government is chosen as the data source investigated. The need for exhaustive examination of such repositories in the CPS Anomaly Detection context is also highlighted. Experimentation is done by applying Tensor Decomposition on the data set after normalization and pre-processing. Details of those phases, as well as the choice of coding platforms, the design of experimental frameworks, timelines estimated, and testing operations, are included in this work. The outcomes are the defined patterns extracted and their analysis-interpretation defended by proofs from actual events of the Project Trial phase.

## TABLE OF CONTENTS

<b>CERTIFICATE OF ORIGINAL AUTHORSHIP</b>	<b>ii</b>
<b>ACKNOWLEDGEMENT</b>	<b>iii</b>
<b>ABSTRACT</b>	<b>iv</b>
<b>TABLE OF CONTENTS</b>	<b>v</b>
<b>LIST OF FIGURES</b>	<b>x</b>
<b>LIST OF TABLES</b>	<b>xiii</b>
<b>Chapter 1 : Introduction</b>	<b>1</b>
1.1 OUTLINE	1
1.2 DOMAIN OF RESEARCH	1
1.2.1 Background	1
1.2.2 Context of Study	1
1.2.3 Problem Domain	2
1.2.4 Scope of Research	2
1.2.5 Important Terminology	3
1.3 RESEARCH QUESTION AND HYPOTHESIS FORMULATION	4
1.3.1 Research Gap	4
1.3.2 Hypothesis and Proposed Solution	4
1.3.3 Functional Advantages	5
1.3.4 Research Objectives	5
1.3.5 General Study Goals	5
1.4 LEARNING OUTCOMES DURING PERIOD OF STUDY	5
1.4.1 Subject Oriented	5
1.4.2 Practice Induced (Apart from technical knowledge)	6
1.5 THESIS ORGANIZATION	6
1.6 CONTRIBUTION IN PUBLICATIONS	7
1.7 CONTRIBUTION IN RESEARCH	7
1.8 CHAPTER SUMMARY	8

<b>Chapter 2 : Evaluative Review – Background</b>	<b>10</b>
2.1 OUTLINE	10
2.2 DOMAIN OF STUDY	10
2.2.1 Background: Cyber-Physical Systems	10
2.2.2 Typical CPS Installations	11
2.2.3 Computational Systemic Intelligence in CPS	11
2.2.4 Technology Scenario 1: The Internet of Things	13
2.2.5 Technology Scenario 2: The Smart Energy Grid	13
2.2.6 Main Challenges in the CPS Domain	14
2.3 PILOT STUDY: SMART ENERGY GRIDS	15
2.3.1 Organization: The Legacy Grid	17
2.3.2 State-of-the-Art: Smart Energy Systems	19
2.3.3 Shortcomings: The Need for a Smarter Grid Ecosystem	22
2.3.4 Technology Gap: Autonomic Intelligence in Smart Grids	27
2.3.5 Literature Survey	28
2.4 CHAPTER SUMMARY	30
 <b>Chapter 3 : Exploratory Review - Research Gap</b>	 <b>32</b>
3.1 OUTLINE	32
3.2 PROBLEM DOMAIN: MULTIWAY BIG DATA IN CYBER-PHYSICAL SYSTEMS	32
3.2.1 Multiway Big Data in CPS	33
3.2.2 Higher Order Big Data Processing	34
3.2.3. Next Phase: Autonomic Middleware for CPS	36
3.3 RESEARCH GAP	39
3.3.1 Need for Higher-Order High Dimensional Pattern Analysis	40
3.3.2 Multimodal Pattern Imaging and Detection	40
3.3.3 Anomaly Detection in Higher Order Data	41
3.3.4 Research Quest: Defining Latent and Active Features of Associations & Dissociations	43
3.3.5 Literature Survey	45
3.4 REFINED RESEARCH POSTULATE	45

3.4.1 Research Problem in Focus	45
3.5 CHAPTER SUMMARY	46
<b>Chapter 4 : Research Hypothesis</b>	<b>48</b>
4.1 OUTLINE	48
4.2 RESEARCH HYPOTHESIS	48
4.2.1 Core Postulate	48
4.2.2 Developed solution	48
4.3 SOLUTION SPACE	49
4.3.1 Knowledge Representation	50
4.3.2 Challenges and Constraints Involved	50
4.3.3 Multiway High-Dimensional Data Processing	51
4.4 INTRODUCTION TO TENSORS	51
4.4.1 Benefits of Tensor Algebra based Data Processing	52
4.5 RESEARCH PROPOSITION	54
4.5.1 Overview	54
4.5.2 Multi-Way Pattern Processing: Behavior-Anomaly Detection	54
4.5.3 Solution Framework	55
4.5.4 Modelling Data as Decomposed Tensor	56
4.5.5 Anomaly or Outlier Patterns	57
4.5.6 Feature Annotation of Anomaly Patterns	57
4.5.7 Anomaly (Feature) Detection condition	58
4.5.8 Application in the context of the Smart Grid Case study	59
4.6 CHAPTER SUMMARY	60
<b>Chapter 5 : Methodology</b>	<b>62</b>
5.1 OUTLINE	62
5.2 TENSOR FACTORIZATION AND DECOMPOSITION	62
5.2.1 Knowledge Discovery: Significance of TF and TD	62
5.2.2 Physical Interpretation	63

5.2.3 Mechanics of Tensor Factorizations	65
5.2.4 PARAFAC or CANDECOMP (Canonical Decomposition) or CP decomposition Model	66
5.2.5 NTF-1 or Non-Negative Tensor Factorization Model	67
5.2.6 NTF-2 Model	67
5.2.7 Nonnegative Tucker Decompositions (NTD)	68
5.2.8 Computational Significance	71
5.3 LITERATURE SURVEY	72
5.3.1 Data Compression	73
5.3.2 Data Clustering	73
5.3.3 Signal Classification	74
5.3.4 Conceptual Parallels	75
5.4 EXPERIMENTATION FRAMEWORK	75
5.4.1 Definition of Behavior and Anomaly patterns	75
5.4.2 Phase I: Data Acquisition	77
5.4.3 Phase II: Pre-Processing	78
5.4.4 Phase III: Tensorization or Data representation	78
5.4.5 Phase IV: Tensor Processing	78
5.4.6 Phase V: Post-processing: Anomaly Detection in Data Tensors	78
5.4.7 Phase VI: Performance Examination and Learning Curve	79
5.4.8 Implementation Environments	79
5.5 CHAPTER SUMMARY	80
<b>Chapter 6 : Experimental Analysis</b>	<b>82</b>
6.1 OUTLINE	82
6.2 EXPLORATORY STUDY OF DATASET	82
6.2.1 Overview	83
6.2.2 Organization	84
6.2.3 Data Collection and Management	85
6.2.4 Raw Data Imported for Experiments	85
6.3 IMPLEMENTATION FRAMEWORK	86



6.3.1 Computational Environment	87
6.3.2 Bash Scripting and Python Virtual Environment	88
6.3.3 Python Toolboxes for Tensor Decomposition Scripting	89
6.4 EXPERIMENT RESULTS AND DISCUSSION	90
6.4.1 Pre-Processing – 1: Acquisition	90
6.4.2 Pre-Processing – 2: Normalization	92
6.4.3 Pre-Processing – 3: Soft Clustering Devices for specific periods	94
6.4.4 Tensor Processing: Tucker Decomposition	97
6.4.5 Discussion	105
6.5 KNOWLEDGE MINING: REFERENCE FEATURE MATCHING	112
6.5.1 Behavioral Observations	112
6.6 CHAPTER SUMMARY	114
 <b>Chapter 7 : Conclusion and Extensions</b>	 <b>116</b>
7.1 OUTLINE	116
7.2 NEXT PHASE OF RESEARCH	116
7.3 FUTURE WORK: REAL - TIME PROTOTYPING	117
7.4 CHALLENGES ENCOUNTERED	118
7.4.1 Pre-Processing	118
7.4.2 Implementation	118
7.5 SUMMARY	119
 <b>Glossary</b>	 <b>121</b>
<b>Bibliography and References</b>	<b>126</b>

## LIST OF FIGURES

Figure 1.1 Overview of the concept-realm of Cyber-Physical Systems, outlining the scope of research	12
Figure 1.2 Mindmap outline of the dissertation	15
Figure 1.3 Mindmap outline of Chapter 1: Introduction	18
Figure 2.1 Global Renewable Power Capacities by the end of 2015; BRICS countries which are Brazil, the Russian Federation, India, China and South Africa	26
Figure 2.2 Energy system challenges and the role of Smart grids in Response	28
Figure 2.3 Basic smart grid ingredients	29
Figure 2.4 Overview of the growth and evolution of Smart Grid Technologies	30
Figure 2.5 Various forms of Big data volumes generated within the SG environment	38
Figure 2.6 Mindmap outline of Chapter 2: Evaluative Review - Thesis Background	40
Figure 3.1 Mindmap Outline of Chapter 3: Exploratory Review - Research Gap	56
Figure 4.1 Significance of Multiway arrays or tensors in Higher order Big Data Processing	61
Figure 4.2 Generalized Interpretation of Higher Order Data Decomposition Operation	62
Figure 4.3 High Level Description of the Research Proposition	64
Figure 4.4 Proposed Knowledge Representation to facilitate Multiway Anomaly Pattern processing	65
Figure 4.5 Anomaly detection using Tensors as Data Structures	66
Figure 4.6 Tucker Decompositions for anomaly detection algorithm in hyperspectral images	67
Figure 4.7 Mindmap outline of Chapter 4: Research Hypothesis	70
Figure 5.1 A single Tensor signal is a linear combination of tensor dictionary atoms, named tensor fibers or tubes, representing a particular information flow; in aggregation they show intermodal correlation	72
Figure 5.2 Data Tensor represented as tensor columns, a T-product of tensor dictionary and tubal sparse coefficients (Red and white labeling - zero and non-zero tubes)	72
Figure 5.3 Physical significance of Higher Order (Multiway) Factorizations of Big Data, as linear combination of basis components	73
Figure 5.4 Nonnegative Tensor Factorization of a rank-1 third-order tensor, employing different visualization modes	73

Figure 5.5 3rd-order PARAFAC (a) Standard, and (b) Alternative forms; set of 3 matrices using a scalar representation	75
Figure 5.6 3rd-order NTF1, Standard form; set of 3 matrices using a scalar representation	76
Figure 5.7 3rd-order NTF2, (a) Standard form; (b) Extended model; set of 3 matrices using a scalar representation	77
Figure 5.8 (a). Overview of Tucker 3 Decomposition model, (b). and significance of the factors	78
Figure 5.9 Summary of the 3 related Tucker Decomposition models	79
Figure 5.10 Role of Tensor Factorizations and Decompositions as Learning Algorithms	80
Figure 5.11 Discovery of Latent Factors in High-dimensional Data Using Tensor Methods [134] and High-throughput candidate phenotype generation via tensor factorization [135]	81
Figure 5.12 Application of Tensor Factorization for EEG Signal Spectral Decomposition for Intermodal Dependencies	82
Figure 5.13 Application of Non-Negative Tucker Decomposition for EEG Signal Classification	83
Figure 5.14 Anomaly detection using Tucker model: 3rd-order data tensors divided into blocks (a) along the largest dimension mode (all clusters), b) along all modes (varying time windows, 1 month or more)	85
Figure 5.15 Mindmap outline of Chapter 5: Methodology	90
Figure 6.1 Flowchart demonstrating the experimental analysis phase and corresponding functions implemented	91
Figure 6.2 Listing of the major subsystems of the SGSC Trial setup and DGDS Installation Overview	92
Figure 6.3 DGDS project timeline diagram	93
Figure 6.4 Composition and structure of DGDS dataset; dataset timelines for each file and Color coding for timelines	95
Figure 6.5 Cluster nodes performance analysis based n Matlab 2011 benchmark	96
Figure 6.6 UML Representation of Preprocessing - I phase	101
Figure 6.7 UML Representation of Preprocessing - II phase	102
Figure 6.8 UML Representation of Preprocessing - III phase	104
Figure 6.9 Device measurement plots for clustered group output_45_50, MARCH_to_MAY_2013	105

Figure 6.10 Device measurement plots for the clustered group output_75_23, MARCH_to_MAY_2013	106
Figure 6.11 Standard tensor compositions investigated in this study, and the elemental pattern frames in the input	107
Figure 6.12 UML Representation of the Tensorization Stage	108
Figure 6.13 Device vs. Time window vs. Timestamp experiment - Input tensor structure and knowledge content	111
Figure 6.14 Cluster vs. Devices vs. Timestamps or Time windows framework - Input tensor structure and knowledge content	112
Figure 6.15 Interpretation of the obtained factors	116
Figure 6.16 Hinton Plot Analysis of factor 0, March to May 2013	117
Figure 6.17 Hinton Plot Analysis of factor 2	118
Figure 6.18 Hinton Plot Analysis of factor 1	119
Figure 6.19 Mindmap outline of Chapter 6: Experimental Analysis	124
Figure 7.1 Mindmap outline of Chapter 7: Conclusions and Extensions	129

## LIST OF TABLES

Table 2.1 Framework adopted for Systematic Structured Review of Relevant Literature	19
Table 2.2 Survey and Review articles that helped the framing and defining of the Research Problem and statements	36
Table 2.3 Autonomicity Inducing Parameters of The Smart Power grid	38
Table 3.1 Survey and Review articles that helped the framing and defining of the Research Contributions	53
Table 4.1 Contextual Anomaly Detection in Smart Grid elements - Survey (C. Alcaraz, L. Cazorla et al. 2015)	68
Table 6.1 Hardware specifications of the 27 HPC Cluster nodes which provided access for computing and analysis	96
Table 6.2 Timeline and constituent data modules of Pre-Processing Stages 1 and 2	100
Table 6.3 Algorithm for Generation of Unique IDs	101
Table 6.4 Algorithm for Data Extraction into Monthly intervals	103
Table 6.5 Algorithm for Statistical Vector based Soft clustering	104
Table 6.6 Interpretations of the dimensions of the Data Tensors designed in the proposed method	107
Table 6.7 Parameter inputs to the Tensorization and Tucker Decomposition scripts	108
Table 6.8 Algorithm for Tucker Decomposition experiments	110
Table 6.9 The reconstruction errors and variations from original tensor in the testing phase, averaged for the datasets between 2012 and 2014; the 0.0 entries were of resolution E-20 or more	120
Table 6.10 The average Decomposition execution rate for all considered groupings across 2012 - 2014	120
Table 6.11 Test Runs for Device vs. Time vs. Time framework to identify the pattern subscale, over 3 month periods from 2013 to 2014, the prime trial duration of SGSC	121